technische universität dortmund

Fakultät Wirtschaftswissenschaften

Problem Set 1.1

Labour Economics, Winter Term 2025/26

Submit by Sunday, 02 November, 22:45h ***on Moodle!***

Learning objectives

- Create and interprete descriptive statistics
- Conduct ordinary least squares (OLS) regressions
- Interpret omitted variables bias (OVB)

Tasks

Get familiar with R. You can find some books in Moodle under *Readings*. In case it is your first time using R we recommend having a look at chapter 1 of *Using R for Introductory Econometrics*, watching one of the many useful YouTube videos, or doing some of the free R exercises on `https://www.codeacademy.com/` or `https://www.datacamp.com/`. Finally, we find AI tools and coding assistants (e.g. ChatGPT, Claude) to be useful in suggesting solutions to many coding problems.

Download the data `ps1_clean_data.Rda` and open it in R Studio. The data contains various variables. $hours$ is the dependent variable indicating the number of hours worked per day. $motivation$ is the intrinsic motivation of individuals for a successful career. This is compared to the average motivated individual where higher values are associated with higher motivation. $education$ displays the number of years spent on education. $wage$ is the wage per hour. $wage\_premium$ is a dummy that indicates whether an individual received a randomly assigned wage increase of 35% or not. This could be because they were drafted into an income support program like the one in Canada discussed in lecture but will be important only in Problem Set 1.2.

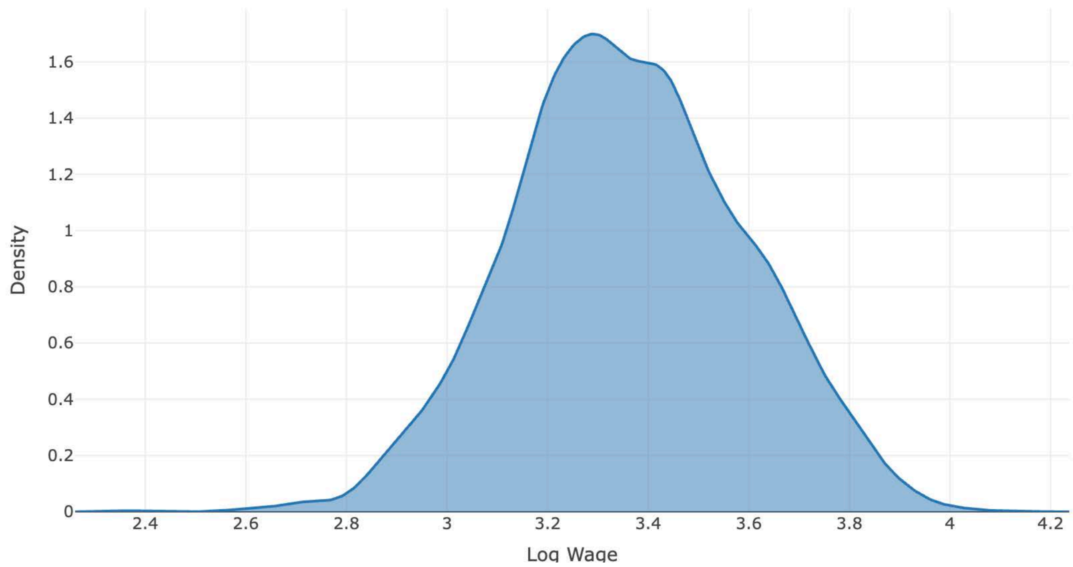
a) Generate the log of wages (ln_wage). Produce a table with descriptive statistics for $education$, $motivation$, $hours$ and ln_wage. Also calculate correlations between these variables and plot the density of log wages as well as a histogram for the years of education. Briefly comment on your results.

   Solution

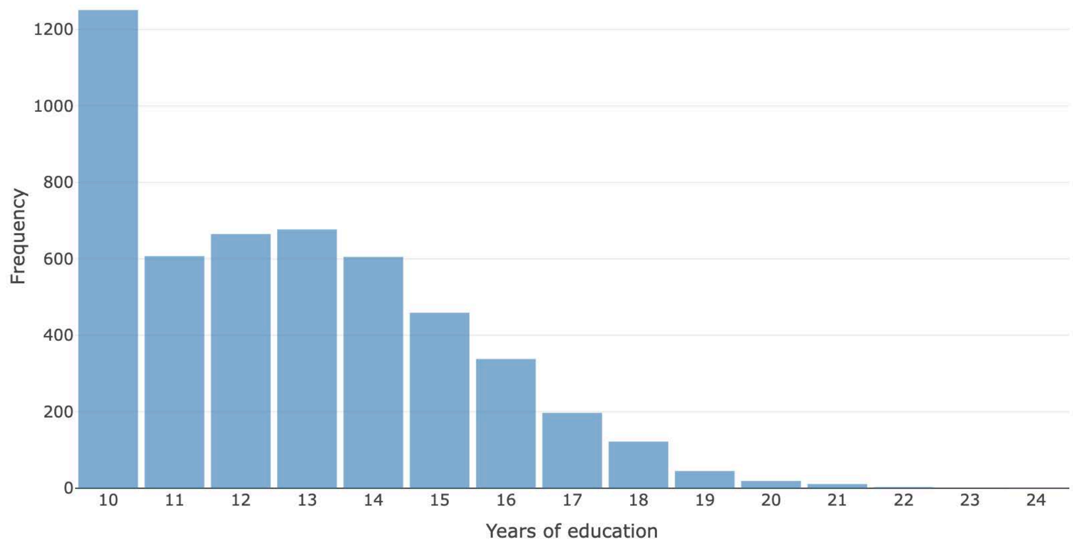
variable	n	min	max	median	iqr	mean	sd	se	ci
<fct>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
education	5000	10	24	12	4	12.8	2.42	0.034	0.067
motivation	5000	-4.56	5.14	-0.002	1.88	0.014	1.39	0.02	0.039
hours	5000	6.35	9.70	7.95	0.666	7.95	0.489	0.007	0.014
ln_wage	5000	2.37	4.12	3.35	0.317	3.36	0.23	0.003	0.006

	education	motivation	hours	ln_wage
education	1.0000	0.1416	0.6463	0.5135
motivation	0.1416	1.0000	0.8155	0.3538
hours	0.6463	0.8155	1.0000	0.6043
ln_wage	0.5135	0.3538	0.6043	1.0000

The summary tells us a lot about the dataframe. We have 5000 individuals; minimum of `education` in our data is 10 years, the maximum 24 years. The median is at 12 years of education, which makes sense as most of the OECD countries have a similar span of education. We can do the same for `hours`. The minimum is 6.35 hours worked, at maximum 9.7 hours. On average, the individuals work 7.95 hours per day, which makes 39.8 hours per week. This is plausible, as once again, most of the developed countries have a weekly workload of 35 hours. So we can think of this data as plausible. All variables are positively correlated with each other. `motivation` and `education` have a relatively weak correlation of 0.138 whereas `education` with `ln_wage` as well as `motivation` with `hours` have relatively strong ones (0.4941 and 0.8045).



The density plot of `ln_wage` looks almost normally distributed with an average of around 3.3 and a range of more than 1 (i.e., 100 log points or $\exp(1) - 1 = 172\%$). We can see a slight second peak at 3.4 which is caused by our wage-premium participants.



The histogram shows us that most people have received 10 years of education (1200 or approx. 24%). After that, only around 600 people received 11 years, with higher numbers for the number of 12, 13 and 14 years. After 14 years, the number of people with higher education declines with a maximum of 24 years. Only few people have a lot of years of education (i.e., master's degree and above).

b) Compare descriptive statistics for individuals who have a *motivation* above or equal to zero versus below. Do the same thing for *education*, 14 years or more (some college) versus below. Again comment on what you find.

Solution

M	variable	n	min	max	median	iqr	mean	sd	se	ci
<dbl>	<fct>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
0	education	2501	10	22	12	4	12.5	2.30	0.046	0.09
0	motivation	2501	-4.56	-0.002	-0.919	1.14	-1.1	0.83	0.017	0.033
0	hours	2501	6.35	9.01	7.62	0.488	7.63	0.364	0.007	0.014
0	ln_wage	2501	2.37	3.90	3.29	0.306	3.30	0.228	0.005	0.009
1	education	2499	10	24	13	4	13.1	2.51	0.05	0.098
1	motivation	2499	0	5.14	0.964	1.19	1.13	0.836	0.017	0.033
1	hours	2499	7.43	9.70	8.24	0.517	8.28	0.373	0.007	0.015
1	ln_wage	2499	2.83	4.12	3.42	0.305	3.42	0.213	0.004	0.008

For the considered people with `motivation` greater than 0, we can find on average higher `education` (13 instead of 12), more `hours` worked (8.28 instead of 7.63) and higher `ln_wage` (3.42 instead of 3.3).

E	variable	n	min	max	median	iqr	mean	sd	se	ci
<dbl>	<fct>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
0	education	3200	10	13	11	2	11.2	1.18	0.021	0.041
0	motivation	3200	-4.29	5.14	-0.118	1.85	-0.1	1.39	0.025	0.048
0	hours	3200	6.35	9.19	7.75	0.558	7.76	0.41	0.007	0.014
0	ln_wage	3200	2.37	3.97	3.27	0.293	3.28	0.213	0.004	0.007
1	education	1800	14	24	15	2	15.5	1.53	0.036	0.071
1	motivation	1800	-4.56	4.17	0.202	1.85	0.216	1.37	0.032	0.063
1	hours	1800	7.08	9.70	8.30	0.558	8.30	0.417	0.01	0.019
1	ln_wage	1800	3.00	4.12	3.47	0.293	3.49	0.196	0.005	0.009

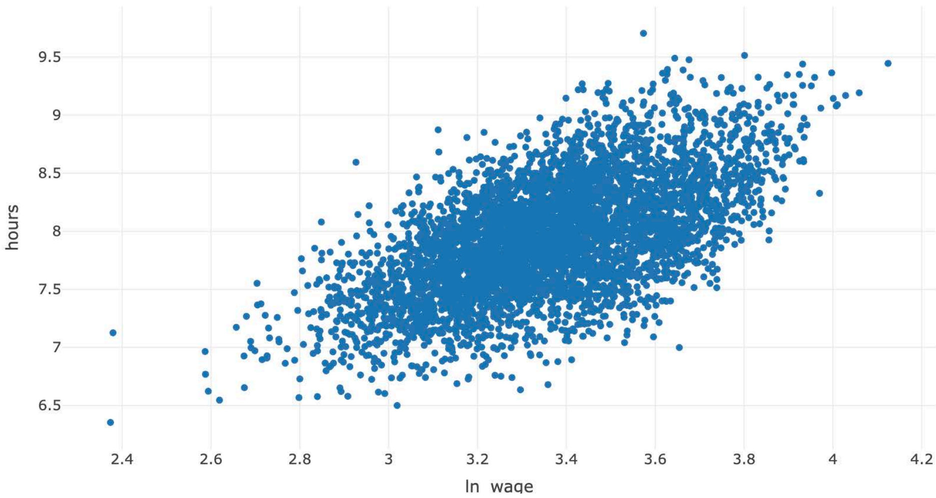
For `education` equal or greater than 14, we can find on average higher `motivation` (0.216 than -0.1), higher `hours` worked (8.3 than 7.76) and higher `ln_wage` (3.49 than 3.28)

So `motivation` and `education` might be important counfounding variables when evaluating the effect of wages on hours worked.

[But we haven't yet tested their economic or statistical significance]

c) Plot a scatter of the hours worked against ln_wage. What do you notice?

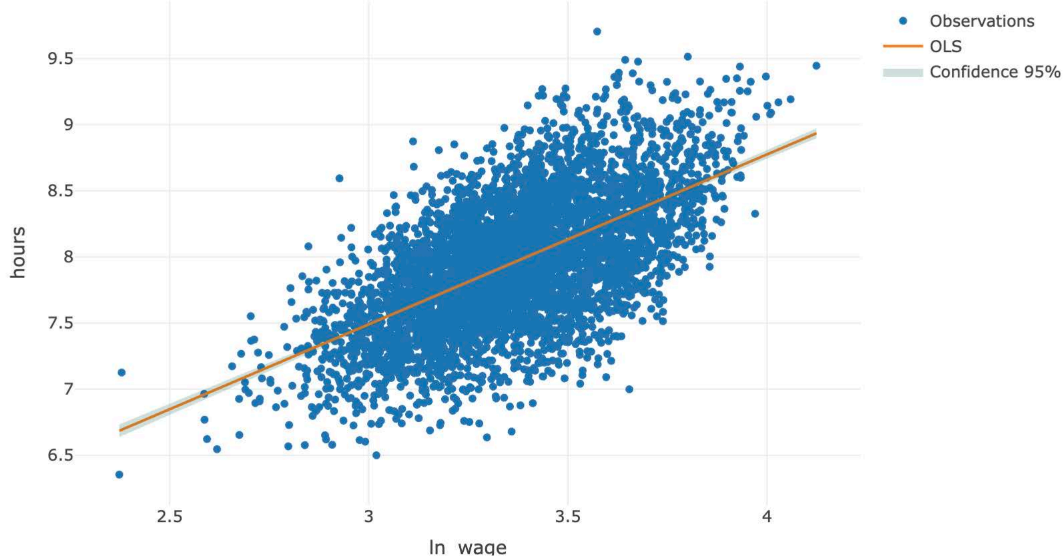
Solution



We see a positive correlation between more `hours` worked and higher `ln_wage`.

d) Do a simple regression of $hours$ on ln_wage. Add the regression line to the plot from c). Include the 95% confidence interval and interpret your results.

Solution




```
Call:
lm(formula = hours ~ ln_wage, data = df1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.33182 -0.27013  0.00022  0.26766  1.47676

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.63582    0.08071   45.05   <2e-16 ***
ln_wage      1.28478    0.02396   53.61   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.39 on 4998 degrees of freedom
Multiple R-squared:  0.3651,    Adjusted R-squared:  0.365
F-statistic:  2875 on 1 and 4998 DF,  p-value: < 2.2e-16
```

The linear regression coefficient is 1.28 which can be interpreted as a semi-elasticity: e.g., a ten percent increase in wages is associated with approximated 0.13 hours (or 7,7 minutes) increase of work per day. We can observe high t-values and p-values close to 0, making the parameter statistically different from zero. Therefore, we have strong evidence for the association between `ln_wage` and `hours`. This overlaps with our analysis of the plot from c).

The confidence band is very tight, which gives us a hint for how precise our estimation really is.

[Still we are careful not to interpret this as a causal effect because, as suggested before, third unobserved variables may affect both wages and hours worked.]

e) Now add *education* to your regression and explain to what extent and why results change.

Solution

```
Call:
lm(formula = hours ~ ln_wage + education, data = df1)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.24877 -0.23175  0.00178  0.23356  1.29310 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.133660   0.071392   57.90   <2e-16 ***
ln_wage     0.786568   0.024323   32.34   <2e-16 ***
education   0.092129   0.002309   39.89   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3397 on 4997 degrees of freedom
Multiple R-squared:  0.5185,	Adjusted R-squared:  0.5183 
F-statistic:  2690 on 2 and 4997 DF,  p-value: < 2.2e-16
```

The coefficient for `education` on hours is positive and statistically significant. Quantitatively, four more years of education (e.g. college vs high school) raise hours by about 5% compared to the average of 8 hours per week.

The coefficient on log wages drops by one third, indeed showing OVB with respect to education. Wages are still significant for hours though.

f) Finally, do the full multivariate regression of *hours* on *ln_wage*, *education*, and *motivation*. Compare your results to before, i.e., out of d)–e), what is your preferred regression specification and results?

Solution

```
Call:
lm(formula = hours ~ ln_wage + education + motivation, data = df1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.36675 -0.06728  0.00121  0.06598  0.33396

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.9681690  0.0223144  267.46   <2e-16 ***
ln_wage     0.2099211  0.0075306   27.88   <2e-16 ***
education   0.0999287  0.0006756  147.91   <2e-16 ***
motivation  0.2499828  0.0010803  231.39   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09925 on 4996 degrees of freedom
Multiple R-squared:  0.9589,    Adjusted R-squared:  0.9589
F-statistic: 3.886e+04 on 3 and 4996 DF,  p-value: < 2.2e-16
```

Adding `motivation` changes the relationship of wages with hours worked again, and strongly so. The coefficient declines from 0.79 to 0.2. That is, a ten percent increase in wages would be associated with only 0.02 hours (or 1.2 minutes) increase of work per day. Small responses of labour supply have been found in the literature, especially among males; we will calculate the exact elasticity in PS 1.2.

As probably expected, all coefficients of hours on `motivation`, `education` and `ln_wage` are positive. They are also statistically significant as indicated by the low p-values (« 0.05). Overall, the three variables (`ln_wage`, `education`, `motivation`) appear to be significantly related to the dependent variable (`hours`).

Given especially the discussion about OVB and our aim to study the "causal effect of wages on labour supply", the preferred specification is the full multivariate regression from g). We should include factors that likely affect individuals' hours and wages at the same time. Controlling in regression for this is the classic way to reduce OVB, and thus we should be closer to the true causal effect. The resulting estimate is also broadly plausible (in line with literature).

R^2 is really high. This is due to this being simulated data; in actual observable data there would be many other factors and R^2 would be lower. For causal analysis, R^2 is secondary anyway, since we want to extract the effect of one specific (policy-relevant) factor; not analyse all factors that drive hours together which is also extremely hard.

Notes: You can work in teams of 1–3 students. Please upload your code as well as a pdf-file with discussions on what you found in the data in response to the tasks above. It should be clear which lines of code and answers in the `.pdf` refer to which question.